УДК 004.855.5

ПРОГНОЗИРОВАНИЕ ВЕРОЯТНОСТИ ВОЗНИКНОВЕНИЯ БРОНХИАЛЬНОЙ АСТМЫ У ДЕТЕЙ С ПРИМЕНЕНИЕМ АЛГОРИТМА СЛУЧАЙНОГО ЛЕСА

Баширов А.Н., Воронов В.И.

ФГБОУ ВО «Московский технический университет связи и информатики», Москва, e-mail: anb888@bk.ru, vorvi@mail.ru

Бронхиальная астма является наиболее частым хроническим заболеванием у детей, в особенности в крупных городах с развитой промышленной базой. В 80% случаев болезнь начинает проявлять себя у детей в возрасте до 6 лет. Достаточно часто эта патология диагностируется некорректно, что приводит к неправильному лечению. Таким образом, разработка различных программных средств, прогнозирующих вероятность формирования бронхиальной астмы, для своевременного диагностирования, становится актуальной и востребованной. В статье, на основе аллергологических данных о 333200 пациентах, разработана модель классификации и прогнозирования вероятности возникновения бронхиальной астмы у детей раннего возраста. Основой для классификации стал алгоритм «случайного леса». Для обучающего набора экспериментальным путём подобраны параметры классификатора. Максимальная точность достигается при 100 деревьях решений. Оптимизация обучающего набора и алгоритма позволила сбалансировать прогнозную эффективность. На стадии обучения точность составила 80%, на тестовой выборке – 79%, при этом степень вероятности принадлежности к классу отдельно взятых точек достигала 85%. «Случайный лес» доказал свою применимость на выбранном наборе данных, однако алгоритм требователен к вычислительным ресурсам.

Ключевые слова: астма, бронхиальная астма, аллергия, дети, диагностика, машинное обучение, искусственный интеллект, анализ данных, алгоритм, модель, случайный лес, бэггинг

PREDICTION OF THE PROBABILITY OF OCCURRENCE OF BRONCHIAL ASTHMA IN CHILDREN WITH THE USE OF THE RANDOM FOREST ALGORITHM

Bashirov A.N., Voronov V.I.

Moscow Technical University of Communications and Informatics (MTUCI), Moscow, e-mail: anb888@bk.ru, vorvi@mail.ru

Bronchial asthma is the most common chronic disease in children, especially in industrial agglomerations. In 80% of cases, the disease begins before the age of 6 years. Often this pathology is diagnosed incorrectly and this leads to improper medical treatment. Therefore, various software tools development for prediction of bronchial asthma formation probability, for timely diagnosis, it is very actual. Based on allergological data of 333200 patients, the article developed a model for classifying and predicting the likelihood of bronchial asthma in young children. The classification was based on the random forest algorithm. For the training set classifier parameters were experimentally selected. Maximum accuracy is achieved with 100 decision trees. Optimization of the training set and algorithm allowed balancing predictive efficiency. At the training stage, the accuracy was 80%, on the test sample – 79%, while the probability of belonging to the class of individual points reached 85%. «Random Forest» proved its applicability with the selected data set, however the algorithm is demanding on computing resources.

Keywords: asthma, bronchial asthma, allergies, children, diagnostics, machine learning, artificial intelligence, data analysis, algorithm, model, random forest, bagging

Медицина является одной из перспективных областей для внедрения искусственного интеллекта. В [1] отмечается, что методы машинного обучения предоставляют широкие возможности в прогнозировании рисков формирования различных заболеваний.

В работе [2] проанализировано применение методов машинного обучения в проблеме детской астмы. Авторы отмечают, что анализ данных, основанный на алгоритмах машинного обучения, облегчает обнаружение скрытых структур в больших данных системы здравоохранения. Проведенные исследования позволили получить важные сведения о механизмах предотвращения и профилактики заболеваний, связанных с астмой.

Методы и алгоритмы машинного обучения также могут применяться в решении самых разных диагностических и реабилитационных медицинских задач. Например, в работе [3] авторы используют параллельные вычисления для высоконагруженных алгоритмов распознавания жестов. Одновременно глубокие нейронные сети находят свое применение в этой же области, пример этого показан в [4]. В работе [5] рассказывается о создании полноценной автоматизированной системы для облегчения социализации людей с нарушениями слуха.

Бронхиальная астма – одно из наиболее частых хронических заболеваний у детей, в особенности в крупных промышленных городах и мегаполисах. Нередко данная патология не диагностируется, и пациенты

лечатся неправильно. В этой связи является актуальной разработка программных средств, с необходимой точностью прогнозирующих вероятность формирования астмы. Пример существующего диагностического алгоритма приведен в [6].

В детской больнице Филадельфии (США) проведено ретроспективное когортное исследование [7] с применением логистической регрессии для выявления зависимостей между видами пищевой аллергии и респираторной аллергией. Объектом исследования выступил набор данных [8], сформированный из сведений о 333 200 пациентах, включающих информацию о наличии различных видов пищевой аллергии, атопического дерматита, аллергического ринита и астмы. В результате была установлена ассоциированность развития астмы и ринита у детей с определёнными видами пищевой аллергии.

В исследовании, описанном в настоящей статье, мы также использовали набор данных [8] для целей проектирования программного модуля на базе алгоритма «случайного леса», прогнозирующего риск формирования бронхиальной астмы при наличии конкретных видов пищевой аллергии, а также аллергического ринита и атопического дерматита.

Цель исследования: выяснение применимости алгоритма «случайного леса» для задачи прогнозирования риска возникновения бронхиальной астмы при наличии конкретных видов пищевой аллергии, а также аллергического ринита и атопического дерматита.

Предобработка данных и параметризация алгоритма «случайного леса»

В работе использовались результаты нескольких исследований отечественных и зарубежных учёных, посвящённые алгоритму «случайного леса».

Корреляционный анализ исследуемого набора данных [8] и его признаков показал, что наиболее сильные линейные связи имеются между астмой и аллергическим ринитом, а также между аллергиями на орехи, молоко и сою (рис. 1).

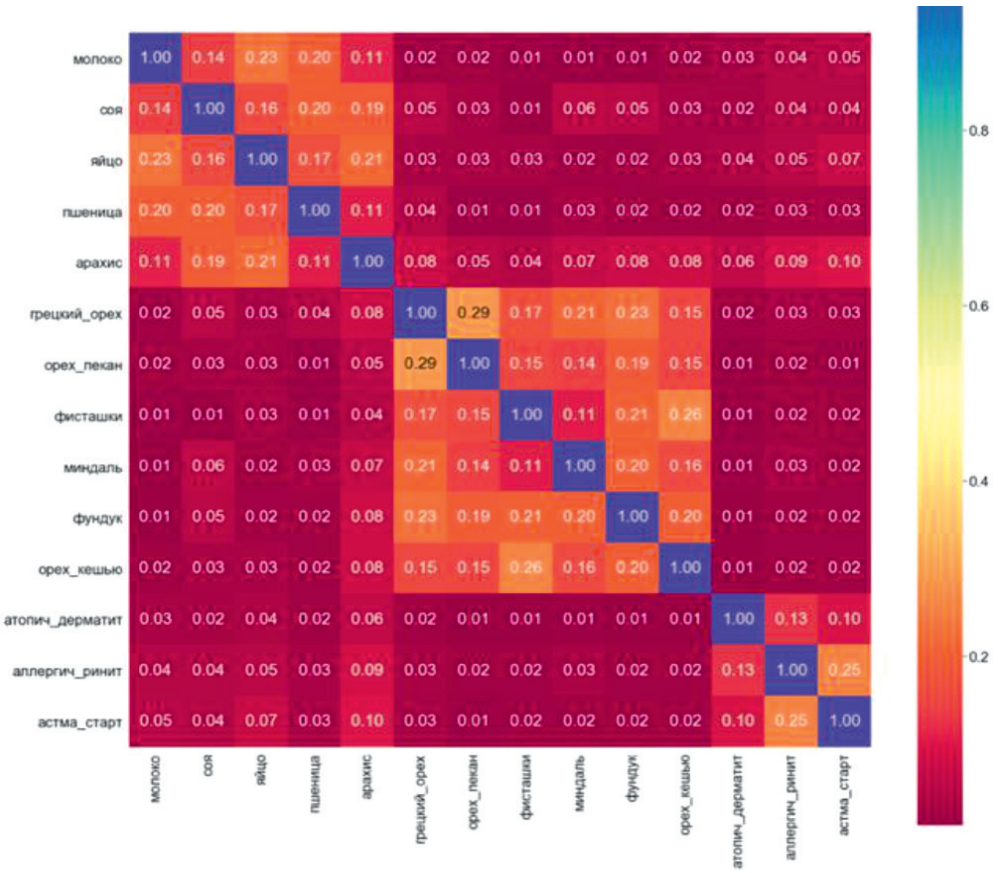

Рис. 1. Матрица коэффициентов связей (14 признаков)

Вместе с тем значения в представленном наборе данных распределены неравномерно, имеются значительные пропуски, что является характерным признаком практически любых медицинских данных. Несмотря на то, что ансамблевые алгоритмы (к которым относится и «случайный лес»), достаточно устойчивы к непропорциональным и разреженным данным, необходимо стремиться к нормализации данных.

Нами из обучающего множества были исключены признаки, не оказывающие влияния на отнесение к классу «имеется риск возникновения бронхиальной астмы» (раса, этническая принадлежность, коэффициент плательщика и сведения о временных эпизодах использования рецептурных медикаментов), а весь набор данных, состоящий из 43 признаков и 330 200 обучающих объектов: год рождения, пол, атопический марш и виды аллергических заболеваний (возраст их начала и окончания) – был масштабирован, приведён к одинаковой числовой шкале, а отсутствующие значения заменены на 0. Для масштабирования данных мы применили модуль «StandardScaler» из библиотеки «Scikit-learn».

Алгоритм «случайный лес» популярен в приложениях машинного обучения по причине его хорошей классификационной способности, универсальности, масштабируемости и достаточно простой программной реализации. «Случайный лес» рассматривается как ансамбль деревьев решений, представляющий собой устойчивую модель, с лучшей ошибкой обобщения и меньшей восприимчивостью к переобучению [9].

Число деревьев – один из основных параметров этого алгоритма. При подборе количества деревьев следует учитывать, что чем их больше, тем эффективнее становится модель на обучающей выборке, но существенно возрастает время обучения, а также растет потребление вычислительных ресурсов.

Другим важным гиперпараметром случайного леса является максимальная глубина деревьев. При его увеличении возрастает качество как на стадии обучения, так и на перекрёстной проверке. Однако в случае очень большого числа объектов в выборке могут получиться крайне глубокие деревья, построение которых занимает значительное время и непозволительно большие ресурсы оперативной памяти.

В библиотеке Scikit-learn реализованы бинарные деревья решений, где каждый родительский узел расщепляется на два дочерних узла. Подобная реализация позволяет уменьшить комбинаторное пространство поиска. При этом в бинарных деревьях решений используются три меры неоднородности (критерия расщепления): Джини, энтропия и ошибка классификации. В Scikit-learn по умолчанию установлена мера неоднородности в форме критерия Джини, который определяет вероятность ошибочной классификации [10].

В родительском узле дерева решений осуществляется классификация данных и переход на левую или правую дочерние вершины. Используя признаки из обучающей выборки, дерево обучается, чтобы сделать выводы о метках классов. Каждое дерево строится по выборке, получаемой из обучающей, с помощью бутстрэпа [10], а именно, из набора данных равномерно берётся n образцов (объектов) с возвращением. При таких условиях непременно окажутся повторы того или иного образца (объекта выборки) со случайной частотой [11].

Для классификации в каждой вершине используется фиксированное число случайно отобранных признаков обучающей выборки и каждый лист дерева содержит данные одной метки класса. Процесс начинается в корне дерева и объект уходит к той дочерней вершине, признак которой дает максимальную информативность. Далее процедура классификации повторяется в итеративном режиме в каждом дочернем узле, пока объект не придет в листовую вершину, где все объекты принадлежат одному и тому же классу [10].

Для классификации и перехода в дочерние узлы определяется целевая функция на самых информативных признаках, которую следует оптимизировать с помощью алгоритма обучения. Целевая функция состоит в максимизации прироста информации при каждой итерации [10]:

$$I\left(d_p, f\right) = i\left(d_p\right) - \sum_{j=1}^{m} \frac{n_j}{n} i\left(d_j\right), \qquad (1)$$

где f – признак, по которому выполняется классификация, d_p и d_j – набор данных родительского и j-го дочернего узла, i – мера неоднородности, n_p – общее количество образцов в родительском узле и n_j – число образцов в j-м дочернем узле. Под приростом информации подразумевается разница между неоднородностью родительского узла и суммой неоднородностей дочерних узлов [10].

Авторы данной статьи реализовали алгоритм случайного леса на языке программирования Python 3.7. В сценарии алгоритма выделено несколько этапов:

1) извлечение случайной бутстрэп-выборки размера *n*;

2) построение деревьев решений из бутстрэп-выборки;

3) в каждом узле дерева: случайным образом отбираются признаки без повторов; строятся дочерние узлы, с использованием отобранного признака, обеспечивающего наилучшее разделение в соответствии с целевой функцией;

4) повторяются первый и второй этапы *k* число раз и на заключительной стадии, для назначения метки класса агрегирован прогноз из каждого дерева решений на основе большинства голосов.

Классификация в методе «случайного леса» осуществляется путём мажоритарного голосования: класс, набравший наибольшее количество голосов деревьев (более 50 %), становится ответом [9].

Результаты исследования и их обсуждение

В соответствии с методикой, нами построен случайный лес, прогнозирующий возможный риск бронхиальной астмы. Первоначально в модели использовалось 60 деревьев решений с максимальной глубиной до 15 вершин. В результате, на стадии обучения коэффициент точности достиг 95 %, а на тестовой выборке – 82 %. Однако степень вероятности принадлежности отдельно взятых экземпляров к классу составила только 74,8 %. При этом в ходе анализа матрицы погрешностей для тестовых прогнозов выявлена недостаточная эффективность классификатора. Исходя из сформированной диаграммы, модель правильно прогнозировала вероятность возникновения астмы в 4349 случаях и ошиблась в 2996 случаях. Вместе с тем в 77862 случаях модель верно определила отсутствие риска астмы и ошиблась в 14753.

Получившиеся результаты свидетельствуют о том, что модель обучилась лучше распознавать отсутствие вероятности возникновения астмы, чем риск её возникновения. Это объясняется тем, что исходный набор данных был асимметричен, т.е. содержал больше данных, относящихся к одному классу. В нем представлено 269326 экземпляров, относящихся к первому классу – «отсутствие вероятности астмы», и только 63874 объекта для второго класса. При обучении построенное дерево решений уделяет повышенное внимание классам с большим число обучающих образцов.

Также, анализируя график обучения, приведенный на рис. 2, можно заметить, что прогнозная эффективность модели не увеличивалась с ростом объёма входных данных.

Результаты работы модели как на стадии тренировки, так и на перекрёстной проверке не изменялись с ростом числа обучающих образцов.

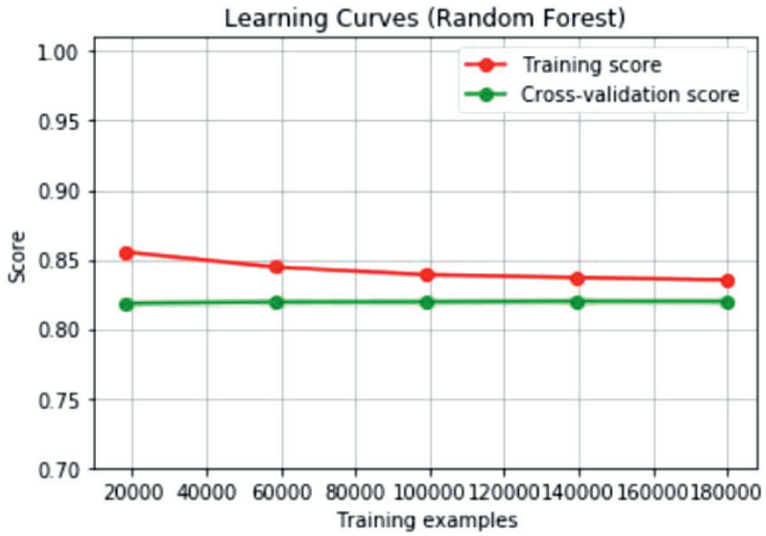

Рис. 2. Кривая обучения (метод случайного леса)

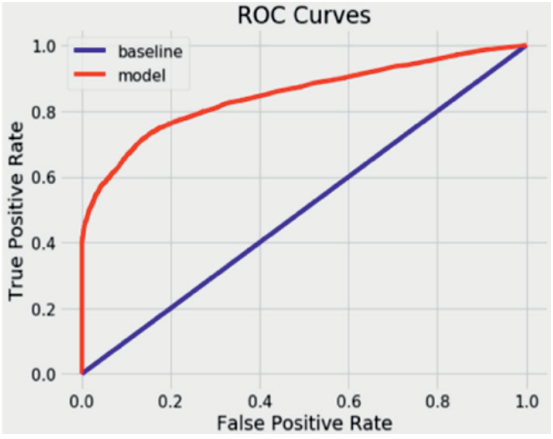

Рис. 3. График ROC-кривой

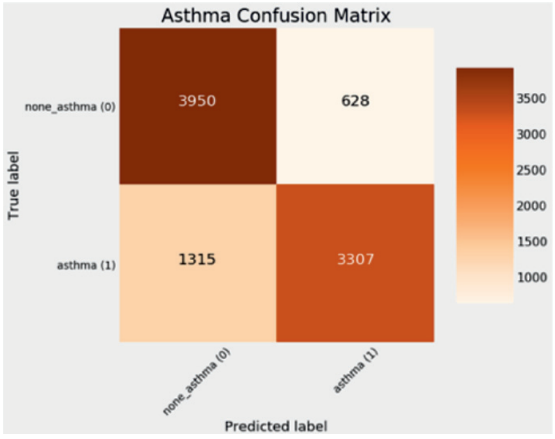

Рис. 4. Матрица погрешностей

В целях оптимизации модели набор данных был сокращён до 44000 образцов, а классы равномерно распределены. Для модифицированного набора, экспериментальным путём, вновь подобраны параметры классификатора. Наиболее высокие показатели эффективности соответствовали 100 деревьям в модели, с максимальной глубиной 25 вершин, с максимальным числом признаков для классификации – 17 и максимальным количеством объектов в вершинах – 15.

Для определения эффективности работы классификатора построен график ROC-кривой и матрица погрешностей для тестовых прогнозов [12].

На графике (рис. 3) ось *x* представляет долю ложных положительных классификаций, ось *y* – долю верных положительных классификаций. Количественную интерпретацию ROC даёт показатель AUC – площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор [11]. В данном случае кривая лучшей модели стремится вверх и вправо.

Согласно приведенным данным (рис. 4), оптимизация обучающей выборки и модели случайного леса позволила сбалансировать прогнозную эффективность. На стадии обучения коэффициент точности составил 80 %, а на перекрёстной и тестовой выборке – 79 %. При этом вероятность принадлежности отдельно взятых экземпляров к классу возросла на 10 % и составила 85 %, то есть доля правильно классифицированных объектов увеличилась.

Анализируя получившиеся деревья решений, одно из которых визуализировано при помощи модуля «GraphViz», можно проследить процесс построения дерева. Для анализа нами выбран фрагмент дерева решений.

На рис. 5 видно, что вероятность возникновения астмы установлена в нескольких случаях, например следуя от элемента «арахис» с мерой неоднородности, равной 3,599, к атопическому дерматиту. Дерево построено до исчерпания выборки, то есть пока в листовых вершинах не остались представители только одного класса.

Для дополнительной проверки спроектированной модели импортированы данные, не встречавшиеся ни в одной из выборок, где смоделированы ситуации для 14 пациентов с наличием некоторых видов пищевой аллергии, а также атопического дерматита и ринита. В результате тестирования алгоритм из имеющихся данных определил вероятность возникновения бронхиальной астмы в 3 случаях, у пациентов: 1-й – 69 %, 2-й – 76 %, 3-й – 55 %. В первом случае у пациента имелась аллергия на молоко, а также атопический дерматит и аллергический ринит, который завершился в возрасте 8 лет. Во втором случае у пациента был зафиксирован аллергический ринит, а у третьего – аллергия на арахис, орех кешью.

Выводы

Авторами построена прогнозная модель риска возникновения бронхиальной астмы у детей в связи с сопутствующими аллергическими реакциями на основе алгоритма случайного леса. Компьютерные эксперименты, проведенные с параметризацией классификаторов и оптимизацией обучающего набора данных, позволили сбалансировать прогнозную эффективность и добиться достаточно высоких результатов в классификации и прогнозировании патологии.

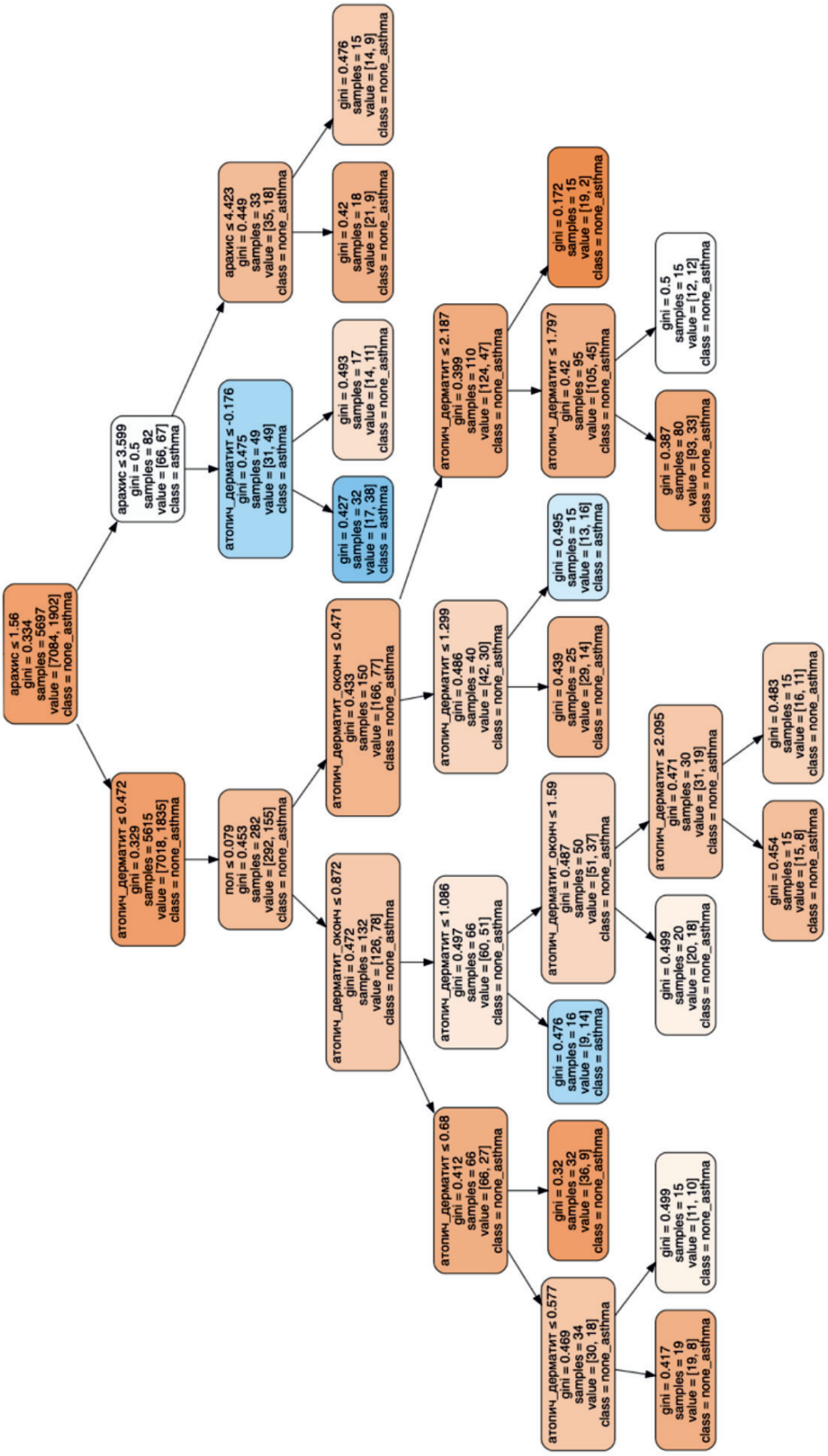

Рис. 5. Пример построения фрагмента случайного леса – один из корневых узлов

Подтверждена применимость метода случайного леса для задач прогнозирования риска возникновения заболеваний при условии использования полноценных ретроспективных когорт данных.

Список литературы

1. Deo R.C. Machine learning in medicine. Circulation, Lippincott Williams & Wilkins. 2015. vol. 132. no. 20. P. 1920–1930.

2. Saglani S., Custovic A. Childhood asthma: advances using machine learning and mechanistic studies. American journal of respiratory and critical care medicine. 2019. vol. 199. no. 4. P. 414–422.

3. Воронов В.И., Воронова Л.И., Генчель К.В. Применение параллельных алгоритмов в нейронной сети для распознавания жестового языка // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2018): VII Международная научно-техническая и научно-методическая конференция: сборник научных статей. В 4-х т. Под ред. С.В. Бачевского. 2018. С. 207–212.

4. Voronov V .I., Genchel K.V., Artemov M.D., Bezumnov D.N. «Surdotelephone» project with convolutional neural network. 2018 Systems of Signals Generating and Processing in the Field of on Board Communications 2018 (Moscow, 14–15 march, 2018). M.: Institute of Electrical and Electronics Engineers Inc., 2018. P. 8350581.

5. Goncharenko A.A., Voronova L.I., Voronov V.I., Ezhov A.A., Goryachev D.V. Automated support system designing for people with limited communication. 2018 Systems of Signals Generating and Processing in the Field of on Board Communications 2018 (Moscow, 14–15 march, 2018). M.: Institute of Electrical and Electronics Engineers Inc., 2018. P. 8350585.

6. Фурман Е.Г., Грымова Н.Н., Санакоева Л.П., Крылова О.А., Мазунина Е.С. Оценка риска развития бронхиальной астмы у детей раннего возраста с помощью опросника «Asthma Prediction Tool» // Российский вестник перинатологии и педиатрии. 2018. № 63(1). С. 34–39.

7. Hill A.D., Grundmeier R.W., Ram G., Spergel J.M. The epidemiologic characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children: a retrospective cohort study. BMC Pediatrics. 2016. DOI: 10.1186/s12887-016-0673-z.

8. Allergy dataset. Zenodo. DOI: 10.5281/zenodo.44529.

9. Пальмов С.В., Денискова А.О. Случайный лес: основные особенности // Наука сегодня: теоретические и практические аспекты: материалы международной научно-практической конференции (Вологда, 28 декабря 2016 г.). Вологда: ООО «Маркер», 2017. С. 51–53.

10. Рашка С. Python и машинное обучение / Пер. с англ. А.В. Логунова. М.: ДМК Пресс. 2017. 418 с.

11. Купенова Э.М., Кашницкий А.В. Метод случайных лесов в задачах классификации спутниковых снимков // Вестник Тверского государственного университета. Серия: география и геоэкология. 2018. № 3. С. 99–107.

12. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006. vol. 27. no. 8. P. 861–874.